PRÁCTICA FINAL

1. Introducción

En este trabajo se estudia un problema de clasificación supervisada multiclase que tiene como objetivo identificar la especie de una flor del género Iris a partir de medidas morfológicas básicas. El sistema recibe como entrada cuatro variables numéricas: longitud y ancho del sépalo, y longitud y ancho del pétalo; y produce como salida la especie predicha entre Setosa, Versicolor y Virginica.

Este tipo de sistema es relevante en el apoyo a la investigación botánica, por ejemplo, donde estudiantes, docentes o investigadores pueden utilizar este modelo para analizar y comparar características morfológicas de distintas especies. En este caso, el stakeholder no solo necesita una predicción correcta, sino también comprender por qué el modelo toma una determinada decisión, especialmente cuando se comparan especies con características similares.

Aunque el conjunto de datos Iris, es un ejemplo sencillo y comúnmente utilizado, el modelo empleado no es directamente interpretable, lo que no nos permite entender qué variables influyen más en las decisiones del sistema. Por este motivo, utilizar la explicabilidad nos permite verificar que el modelo se basa en patrones coherentes, y aumenta la confianza del usuario en el sistema.

2. Datos y métodos

2.1. Dataset

Se ha utilizado el dataset Iris, que es un dataset público disponible académica y es comúnmente utilizado ya que está integrado directamente en bibliotecas como *scikit-learn*. El dataset contiene 150 entradas, cada una descrita mediante cuatro variables numéricas continuas: longitud y ancho del sépalo, y longitud y ancho del pétalo. Los datos se distribuyen de forma equilibrada entre las tres clases (50 muestras de cada clase) correspondientes a las especies Setosa, Versicolor y Virginica. Esto evita problemas derivados del desbalanceo de clases.

Los datos son tan completos y sencillos que no necesitan ningún método para tratar valores erróneos o falta de ellos, ni requieren de un proceso complejo de preprocesado.

2.2. Modelo y evaluación

El modelo que se ha entrenado es un Random Forest compuesto por 200 árboles de decisión, lo que permite capturar relaciones no lineales complejas entre las variables de entrada. El modelo tiene un buen rendimiento predictivo y una alta estabilidad, aunque al tener tantos árboles de decisión no es directamente interpretable.

El dataset se ha dividido en un 70% para el entrenamiento y un 30% para la evaluación, con estratificación por clase para que los datos tuvieran la distribución original de las especies. Esto se ha podido hacer así ya que estamos ante datos tabulares no temporales y así podemos evaluar la generalización del modelo de forma fiable. El rendimiento del modelo se ha evaluado con la métrica de accuracy, dado que es un problema de clasificación multiclase con clases balanceadas.

Los resultados obtenidos muestran un accuracy del 100% en el entrenamiento y de un 91.1% en evaluación. Esta pequeña diferencia entra dentro de la normalidad y nos muestra que el modelo no está sobreajustado. Estos valores tan altos indican que el modelo generaliza bastante bien a datos que no ha visto antes.

2.3. Técnicas de explicabilidad empleadas

Para analizar el comportamiento del modelo, se han empleado distintas técnicas de XAI tanto a nivel global como local.

A nivel global, se ha utilizado SHAP para identificar la importancia de las variables en el dataset, teniendo en cuenta el carácter multiclase del problema. Es decir, se ha aplicado SHAP para las tres clases de manera separada de forma que se puede analizar el efecto de las variables y su importancia para cada clase. Además, se ha utilizado FIP (Feature Importance Permutation) para evaluar la relevancia global de cada característica midiendo la caída del rendimiento del modelo al permutar sus valores. Con este método se ha querido complementar las explicaciones de SHAP, además, de servir como validación adicional, permitiendo comprobar si las variables identificadas como importantes por SHAP también tienen un impacto significativo en la capacidad predictiva del modelo.

A nivel local, se han aplicado las técnicas de SHAP local y LIME para explicar las predicciones individuales. Se ha escogido SHAP local ya que permite cuantificar la contribución de cada variable a la predicción concreta, mostrando cómo cada característica empuja la decisión del modelo a favor o en contra de la clase predicha. Por otra parte, se ha decidido complementar con LIME que proporciona una explicación alternativa basada en la aproximación local del modelo mediante un clasificador interpretable. De esta manera se

puede comparar entre distintos métodos de explicabilidad y tener una idea más completa de lo que le importa al modelo y lo que no.

Por último, se ha implementado un sanity check adicional tanto a nivel global como local mediante las técnicas de feature ablation y perturbación de variables, respectivamente. A nivel global, se ha evaluado el impacto en la accuracy del modelo al eliminar individualmente características relevantes, mientras que a nivel local se ha analizado cómo varían las probabilidades de predicción al modificar valores de las variables más influyentes para un input concreto. De esta manera podemos comprobar la fiabilidad, coherencia y robustez de las explicaciones.

3. Resultados

3.1. Rendimiento del modelo

El modelo entrenado tiene muy buen rendimiento como ya se ha mencionado anteriormente, con un accuracy de entrenamiento del 100% y un accuracy de evaluación del 91.1%. En la Figura 1 se puede apreciar la matriz de confusión donde se muestra un comportamiento casi perfecto para la clase Setosa. Los errores se concentran principalmente entre Versicolor y Virginica, lo cual es coherente con la literatura del dataset Iris, ya que ambas especies tienen solapamientos morfológicos.

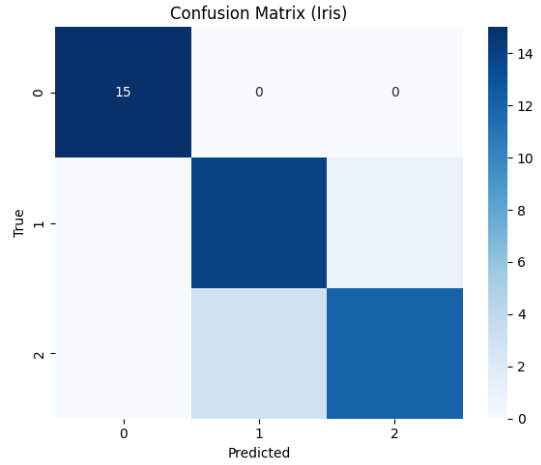

Figura 1: Matriz de confusión resultante del modelo

3.2. Explicaciones globales

SHAP global (multiclase)

Los resultados muestran patrones claros y coherentes.

- Setosa: en la Figura 2 se puede ver como la longitud y el ancho del pétalo presentan valores SHAP fuertemente negativos cuando son bajos, empujando la predicción hacia

esta clase. Esto concuerda con el conocimiento botánico: Setosa se caracteriza por pétalos cortos y estrechos.

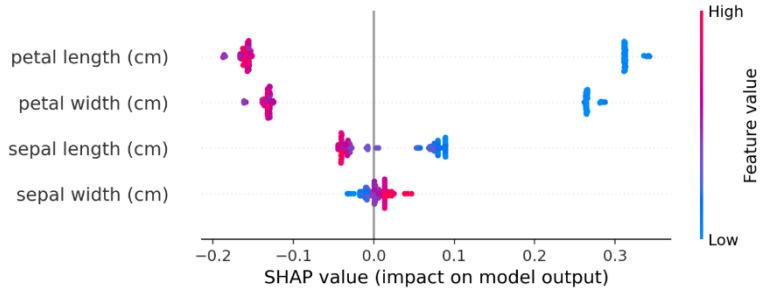

Figura 2: valores SHAP para la clase Setosa

- Versicolor: en la Figura 3 se puede ver como las variables del pétalo siguen siendo las más relevantes, pero con las relaciones entre el valor de la variable y la salida menos claros. La longitud del sépalo adquiere cierta importancia, reflejando la naturaleza intermedia de esta especie.

- Virginica: en la Figura 4 se puede ver como los valores altos de longitud y ancho de pétalo tienen un impacto positivo claro en la probabilidad de esta clase, mientras que valores bajos empujan la predicción hacia otras especies.

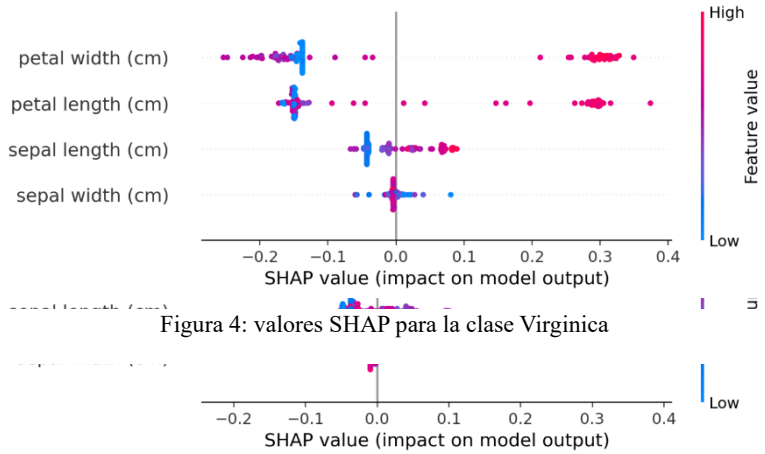

Figura 4: valores SHAP para la clase Virginica

Figura 3: valores SHAP para la clase Versicolor

En todos los casos, las variables del sépalo muestran una contribución menor, lo que indica que el modelo basa sus decisiones principalmente en las medidas del pétalo.

Permutation Feature Importance

La PFI calculada sobre el conjunto de test, refuerza las conclusiones obtenidas en SHAP tal y como se muestra en la Figura 5. La longitud del pétalo es la variable más crítica: al permutarla, la accuracy cae de forma significativa, seguida del ancho del pétalo. Las variables del sépalo tienen un impacto muy reducido sobre el rendimiento global. Además

de proporcionar información de manera independiente, este resultado actúa como una validación de SHAP, ya que ambos métodos identifican el mismo orden de importancia de las características.

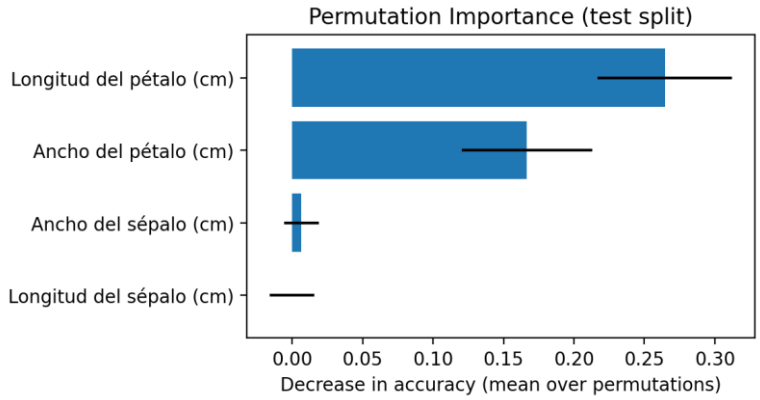

Figura 5: resultado de PFI en el valor del accuracy

Sanity Check: Ablación de variables

Para verificar que las explicaciones globales son realmente informativas, se ha aplicado el método de ablación de variables con validación cruzada. Los resultados mostrados en la Tabla 1 muestran que eliminar una sola variable del pétalo produce una caída en el accuracy moderada, pero que eliminando tanto la longitud como el ancho del pétalo se provoca una caída drástica de la accuracy media. Además, se puede ver como eliminar variables del sépalo apenas afecta al rendimiento.

	features_removed	cv_accuracy_mean	cv_accuracy_std	delta_vs_base
5	petal length + petal width	0.7067	0.0533	-0.2533
4	petal width (cm)	0.9133	0.0718	-0.0467
3	petal length (cm)	0.94	0.0442	-0.02
2	sepal width (cm)	0.9533	0.034	-0.0067
1	sepal length (cm)	0.9533	0.034	-0.0067
0	(none)	0.96	0.0389	0

Tabla 1: resultados del sanity check global con ablación de variables

Esto confirma que las variables identificadas como importantes por SHAP y por PFI son realmente necesarias para la generalización del modelo.

3.3. Explicaciones locales

Para analizar los resultados de las explicaciones locales se ha utilizado la predicción con los valores de las variables predefinidos, por lo que el input ha sido el mostrado en la Tabla 2.

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5	8.7	4	5.8

Tabla 2: valores del input predeterminado

SHAP local

Las explicaciones SHAP locales permiten descomponer la predicción individual en contribuciones por variable. En la muestra utilizada (ver Figura 6) se observa que la longitud del pétalo tiene la contribución positiva más alta hacia la clase predicha y que el ancho del pétalo puede actuar tanto a favor como en contra dependiendo de su valor concreto. Además, se ve como las variables del sépalo tienen un impacto reducido (cercano a cero).

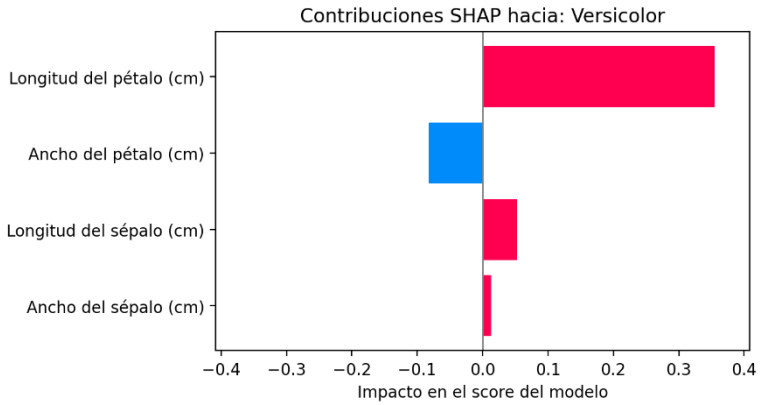

Figura 6: valores SHAP para el input predeterminado

LIME local

LIME proporciona una explicación alternativa basada en un modelo lineal local. Los resultados mostrados en la Figura 7 son coherentes con SHAP. Las reglas más influyentes involucran umbrales sobre longitud y ancho del pétalo, mientras que las contribuciones del sépalo aparecen con pesos bajos.

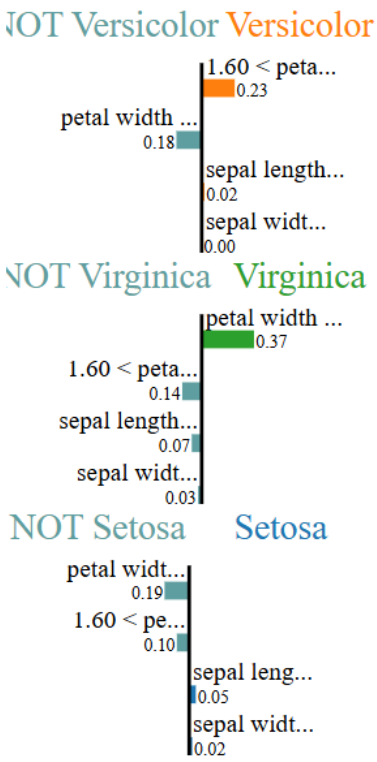

Figura 7: resultados LIME para el input predeterminado

Sanity check local: perturbaciones

Se ha implementado un sanity check local perturbando las variables más relevantes según SHAP. En primer lugar, se han aplicado perturbaciones suaves sustituyendo los valores originales por la mediana del dataset. En este caso, los cambios observados (ver Tabla 3) en la probabilidad son mayores en el ancho del pétalo donde la mediana se encuentra más alejada del valor establecido.

	feature_perturbed	mode	shap_value	original_value	perturbed_value	p_class_original	p_class_after	delta_p	abs_delta_p
0	petal length (cm)	median	0.3552	4	4.35	0.675	0.675	0	0
1	petal width (cm)	median	-0.0822	5.8	1.3	0.675	0.94	0.265	0.265
2	sepal length (cm)	median	0.0524	5	5.8	0.675	0.635	-0.04	0.04
3	sepal width (cm)	median	0.0133	8.7	3	0.675	0.585	-0.09	0.09

Tabla 3: sanity check local con el perturbando con la mediana

Por otro lado, se ha implementado un sanity check más agresivo, en el que las variables se perturban hacia valores alejados de su contribución SHAP. En concreto, si la contribución SHAP es positiva, su valor se desplaza hacia un percentil bajo del dataset (q10), mientras que, si la contribución es negativa, se desplaza hacia un percentil alto (q90). En este otro caso, se observa en la Tabla 4 lo mismo que en la variante anterior. Si el valor perurbado se aleja mucho del valor establecido, como es el caso de la longitud del pétalo, la caída es mayor. Además, se muestra como un cambio proporcional en una variable importante y en una con menos relevancia, tienen diferente impacto. Por ejemplo, en la Tabla 4 se puede ver

cómo tanto el cambio la longitud del pétalo, como el cambio del ancho del sépalo son bastante grandes y el efecto en el modelo es mucho mayor con el cambio de la longitud del pétalo. Esto muestra que es una variable mucho más relevante, tal y como muestran SHAP y LIME.

	feature_perturbed	mode	shap_value	original_value	perturbed_value	p_class_original	p_class_after	delta_p	abs_delta_p
0	petal length (cm)	aggressive	0.3552	4	1.4	0.675	0.33	-0.345	0.345
1	petal width (cm)	aggressive	-0.0822	5.8	2.2	0.675	0.675	0	0
2	sepal length (cm)	aggressive	0.0524	5	4.8	0.675	0.675	0	0
3	sepal width (cm)	aggressive	0.0133	8.7	2.5	0.675	0.61	-0.065	0.065

Tabla 4: sanity check local perturbando con los percentiles (q10 y q90)

El comportamiento de ambas variantes muestra como cambios pequeños no afectan al modelo indicando robustez frente a pequeñas variaciones en las características de entrada, mientras que grandes cambios en las variables con mayor SHAP provocan cambios más relevantes en la salida del modelo.

4. Acciones e insights derivados de las explicaciones

4.1. Limitaciones e interpretaciones potencialmente engañosas

Una primera limitación importante es que no se ha analizado explícitamente la relación entre las variables de entrada. Las explicaciones obtenidas mediante SHAP y FPI evalúan la contribución individual de cada característica, pero no permiten estudiar de forma directa posibles interacciones entre las variables. En el caso del dataset utilizado es razonable pensar que ciertas combinaciones entre medidas del pétalo y del sépalo pueden ser relevantes para distinguir especies concretas, especialmente en regiones donde las clases se solapan.

Relacionado con esto, las explicaciones globales muestran que las variables del pétalo dominan el proceso de decisión del modelo. Sin embargo, esto no implica necesariamente que las variables del sépalo sean irrelevantes en todos los casos. Un análisis basado únicamente en importancias marginales podría llevar a la conclusión errónea de que la información del sépalo puede eliminarse sin consecuencias y así simplificar el modelo. Sin embargo, esta información podría ser útil para diferenciar especies concretas o para mejorar la robuste del modelo en determinadas regiones del espacio de características.

Además, las explicaciones locales pueden resultar estables o inestables dependiendo de la instancia analizada. En predicciones realizadas con alta confianza, las perturbaciones suaves apenas modifican la probabilidad, lo que podría interpretarse erróneamente como una falta de relevancia de las variables. Sin embargo, al forzar cambios más drásticos, la predicción si se ve afectada.

4.2. Acciones y recomendaciones

A partir de las explicaciones obtenidas, se pueden proponer algunas acciones y líneas de trabajo futuras. Por un lado, los resultados sugieren explorar ablaciones por grupos de variables. Este tipo de análisis permitiría responder preguntas relevantes como si es posible distinguir correctamente todas las especies utilizando solo características del pétalo. Este enfoque también podría ayudar a evaluar la redundancia entre variables y a simplificar el modelo si fuera posible.

Por otro lado, desde la perspectiva de stakeholder, es importante destacar que las explicaciones deben utilizarse como herramientas de apoyo al análisis, y no como conclusiones definitivas. Los resultados obtenidos permiten generar hipótesis sobre el comportamiento del modelo, pero estas deben contrastarse mediante experimentos adicionales o con conocimiento del dominio. La combinación de múltiples técnicas de explicabilidad junto con sanity checks, son clave para reducir el riesgo de interpretaciones erróneas y así aumentar la confianza en las conclusiones obtenidas.

5. Discusión

Este trabajo analiza el comportamiento de un modelo de clasificación mediante técnicas de explicabilidad global y local, aunque presenta algunas limitaciones. En concreto, se ha utilizado un dataset pequeño y bien estructurado, lo que facilita tanto el aprendizaje del modelo como la interpretación de las explicaciones, pero limita la generalización de los resultados a escenarios reales más complejos y con ruido.

Además, aunque se han aplicado sanity checks globales y locales, no se ha estudiado explícitamente la interacción entre variables ni la sensibilidad de las explicaciones frente a cambios en el modelo y en sus hiperparámetros. Desde el punto de vista de los riesgos, las explicaciones pueden interpretarse erróneamente como relaciones causales si no se acompañan de conocimiento del dominio.

Como trabajo futuro, se propone extender el análisis a otros modelos, estudiar interacciones entre variables y aplicar estas técnicas a datasets más realistas, así como incorporar métricas cuantitativas para evaluar la estabilidad y fidelidad de las explicaciones.